DOCUMENT RESUME

ED 348 384 TM 018 767

AUTHOR Shermis, Mark D.; And Others
TITLE Assessing the Reliability of Computer Adaptive Testing Branching Algorithms Using HyperCAT.
PUB DATE Apr 92
NOTE 32p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Algorithms; Bayesian Statistics; Comparative Analysis; *Computer Assisted Testing; Estimation (Mathematics); Higher Education; Mathematical Models; Maximum Likelihood Statistics; Test Bias; Test Construction; *Test Reliability; *Undergraduate Students
IDENTIFIERS Ability Estimates; *Branching Algorithms; *HyperCAT Computer Program; Paper and Pencil Tests

ABSTRACT

The reliability of four branching algorithms commonly used in computer adaptive testing (CAT) was examined. These algorithms were: (1) maximum likelihood (MLE); (2) Bayesian; (3) modal Bayesian; and (4) crossover. Sixty-eight undergraduate college students were randomly assigned to one of the four conditions using the HyperCard-based CAT program, HyperCAT. As a way to control for order effects, half of the students were randomly assigned to take the paper-and-pencil test first, followed 3 weeks later by the CAT, while the other half took the CAT first. Investigative analyses showed no initial group differences by algorithm for the paper-and-pencil test and for CAT-estimated ability. In addition, there was no order effect. The internal consistency coefficient for the paper-and-pencil test was 0.73. The marginal reliability for the CAT was 0.97. Correlations between the paper-and-pencil scores and theta estimates of ability ranged from 0.48 to 0.79. Reliability was highest for the MLE algorithm, followed by the Bayesian, modal Bayesian, and crossover algorithms, respectively. Given the constraints of MLE branching algorithms (e.g., the examinee must get at least one item correct and one item incorrect), and the alleged biasedness associated with Bayesian branching strategies, the results suggest that the modal Bayesian testing may provide an acceptable alternative. Six tables present study data. Three figures and 14 references are included. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED348384

Assessing the Reliability of Computer Adaptive Testing Branching Algorithms Using HyperCAT™

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
MARK D. SHERMIS

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

by

Mark D. Shermis

Don McCormack

Shirin Catterson

Department of Educational Psychology

The University of Texas at Austin

Paper presented at the Annual Meetings of the National Council for Educational Measurement, San Francisco, CA, April 20-24, 1992.

Inquiries should be addressed to Mark D. Shermis, Department of Educational Psychology, EDB 504, The University of Texas at Austin, Austin, TX 78712-1296, (Voice) 512-471-4155/(Fax) 512-471-4607/(BITNET) EPFG745@UTXVM/(INTERNET) EPFG745@UTXVM.CC.UTEXAS.EDU

TM018717

BEST COPY AVAILABLE

ABSTRACT

This study examined the reliability of four branching algorithms commonly used in computer adaptive testing (CAT). These algorithms included Maximum Likelihood (MLE), Bayesian, Modal Bayesian, and Crossover. Sixty-eight undergraduate college students were randomly assigned into one of the four conditions using the HyperCard™-based CAT program HyperCAT™. As a way to control for order effects, half the students were randomly assigned to take the paper-and-pencil test first, followed three weeks later by the CAT while the other half took the CAT first. Investigative analyses showed no initial group differences by algorithm for the paper-and-pencil test [$F(3,59) = 0.33$, $p > .05$] and for CAT-estimated ability [$F(3,64) = 1.64$, $p > .05$)]. In addition, there was no order effect [$\lambda = .944$, $F(2,60) = 1.79$, $p > .05$)]. The internal consistency coefficient for the paper-and-pencil test was calculated at $\alpha = .73$. The marginal reliability calculated for the CAT was $\overline{\rho} = .97$. Correlations between the paper-and-pencil score and theta estimates of ability ranged from $r = .48$ to $.79$. Reliability was highest for the MLE algorithm followed by Bayesian, modal Bayesian, and Crossover, respectively. Given the constraints of MLE branching algorithms (e.g., the examinee must get at least one item correct and incorrect) and the alleged biasedness associated with Bayesian branching strategies, the results suggest that the modal Bayesian testing may provide an acceptable alternative.

Assessing the Reliability of Computer Adaptive Testing Branching Algorithms Using HyperCAT™

Introduction

This study examined the reliability of four branching algorithms commonly used in computer adaptive testing (CAT). These algorithms included Maximum Likelihood (MLE), Bayesian, Modal Bayesian, and Crossover. The vehicle for implementing the computer adaptive test was HyperCAT™ which is based on Apple Computer's HyperCard™ product.

Computer Adaptive Testing

Computer adaptive testing usually works in the following way. The computer first determines an initial testing level[1], then randomly selects an item from the item bank that represents that level. According to the responses provided by the student, the computer branches to a more appropriate ability level as determined by latent-trait calibration. Two decision models are typically used for branching: (1) the Bayesian approach (Novick, 1969; Owen, 1969) and (2) maximum likelihood algorithm (Urry, 1977; Reckase, 1973;1974). Work by Ho and Hsu (1989) suggests that two alternative strategies might also be employed: (1) a crossover technique that begins testing with the Bayesian strategy and then switches over to the MLE branching strategy, and (2) a modal Bayesian branching strategy (Mislevy & Bock, 1982).

Most adaptive branching strategies will calculate an individual's ability estimate (θ) based on the characteristics of one, two, or three item parameters: discrimination (a), difficulty (b), and (pseudo-) guessing (c). The three-parameter model calculates the probability of an individual obtaining a θ score by solving the following equation:

---

[1]An initial testing level might be determined after asking a few screening questions.

$$P_i = P(\theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta_i - b_i)}} \qquad [1]$$

where

$P(\theta)$ is the probability that an individual at ability θ will get the item correct,

θ_i = ability for examinee i,

a_i is the item's discrimination,

b_i is the item's difficulty, and

c_i is the (pseudo-) guessing parameter[2].

This model is known as the three-parameter logistic model (3PL) and is based on the assumption that the item pool measures a unidimensional content domain. The model uses a logistic curve rather than the normal curve to simplify some of the mathematical operations used in calculating the ability estimate and its standard error. In practical terms, the logistic curve closely approximates distributions that are normal (e.g., when the constant in Equation 1 is equal to 1.7).

One drawback of using the 3PL model is that it requires a relatively large item pool (e.g., 80 items) and a modest sample size (n = 1000) to reliably calibrate the three item parameters. This makes the 3PL model difficult to implement for classroom or small-scale testing procedures where item pools and sample sizes tend to be smaller. Moreover, many statewide item banks are calibrated for the one- and two-parameter logistic models (1PL and 2PL).

---

[2]Some authors refer to the *c* parameter as the *pseudo*-guessing parameter (cf. Hambleton & Swaminathan, 1985) because even though the probability of getting an item correct can be calculated *a priori* (e.g., five response alternatives = .20 probability of a correct response via random guessing), empirical results seldom match their theoretical predictions.

HyperCard

HyperCard™ is a database programming environment that allows the test administrator to create educational "stackware." Each stack contains "cards" on which objects appear. These objects represent "events" that the author wants the computer to produce. One such "event" might be the presentation of a test item. Figure 1 illustrates what a test item might look like on a HyperCard™ screen. Note that next to each response alternative is a button. The examinee indicates his/her response by navigating to the button corresponding to the desired answer and "clicking." This is similar to how a mouse is used in other computer operations.

Insert Figure 1 About Here

The main advantages that HyperCard™ has over other database environments are its (1) ease of use and (2) ability to customize the environment through user-written "add-in" modules. Most individuals can become active users of HyperCard™ with about an hour's worth of training. In addition, it is not necessary to become a programmer to effectively use the package.

HyperCAT™

HyperCAT™ is a computer adaptive testing system that employs HyperCard™ as a front-end processor. HyperCAT™ uses item response theory (IRT) as the foundation for the adaptive strategies and is currently based on the one-, two-, and three-parameter logistic models. The HyperCAT™ branching strategies are implemented as HyperCard™ XCMDs. XCMDs are user-created "add-in" modules which add to the functionality

of HyperCard™. In this case HyperCAT™'s XCMDs, written in the C programming language, perform all the numerical calculations used to estimate ability and information levels. They are also the mechanism for implementing branching strategies by calculating which item to present next. The compiled C code executes much more quickly than HyperCard™'s native language HyperTalk.

HyperCAT™ is divided into three card sections: (1) Card Manager, (2) Test Administration, and (3) Help. The Card Manager card supports a number of functions, including the provision of information on (a) the current theta or ability level, (b) a standard error for the estimate of ability, (c) a register of the number of items, (d) fields in which to enter information on the difficulty, discrimination, and (pseudo-) guessing parameters of the three-parameter logistic model, (e) user-modifiable fields (initial theta level, maximum number of items to be administered, maximum/minimum ability bounds), (f) a "debug" field to monitor the progress of testing, and (g) utilities to simulate a test, sort cards, print the contents of the "debug" field, and obtain help. Figure 2 shows the layout of the Card Manager in the current configuration of HyperCAT™.

Insert Figure 2 About Here

Test administration generally consists of three steps: a registration procedure, actual administration of the test, and report generation.

Creating a test consists of 4 component activities: (1) calibrating test items, (2), placing the items on cards, (3) developing

buttons for the response alternatives, and (4) implementing a registration procedure.

HyperCAT™ does not currently support a subsystem for item calibration, although one may be eventually developed for the one parameter logistic (Rasch) model. HyperCAT™ can directly import the parameters generated by most of the popular calibration programs if they are output in a standard ASCII text file.

Placing items on the cards can occur through a variety of inputs. Most of the time the item will be created from scratch using the Tool palette available in HyperCard™. Parts of items may be placed in fields (e.g., item numbers) while other components (e.g., item stem and response alternatives) can be accompanied by foreground graphics. Another way to create items is by importing them though the Clipboard or Scrapbook. For example, you may have created the item using a paint program from a third-party vendor, copied the item to the Scrapbook, and placed it in HyperCard™ through the Clipboard. Alternatively you may have an already existing text version of the test in a word processing file. You can simply cut-and-paste items from the text file directly into HyperCard™ through the Scrapbook or Clipboard. Finally you might import the item into HyperCard™ through a scanning device. For example, using Apple's 300dpi OneScan™, you could scan the paper-and-pencil test item(s) directly into HyperCard™ for use in the HyperCAT™ program. The development of response alternative buttons and registration procedures is covered more fully in Shermis (1991).

There are presently two reports that can be generated from HyperCAT™. One of the reports may be printed by the student or test administrator when a student completes the test. Tests are terminated

under "normal" and "error" conditions. Normal test termination occurs when the test is completed with an ability estimate at the information level (within the standard error) requested. An error condition occurs when the examinee has exceeded the maximum number of items specified, the range of θ permitted, or if a requirement of ability-level calculation has not been achieved (e.g., a right/wrong combination for branching based on maximum likelihood).

The second report is located on the Card Manager and is generally reserved for debugging or research purposes. This report tracks all the internal calculations made by the XCMD and the choices made by the examinee.

HyperCAT™ Branching Algorithms

Currently, HyperCAT™ employs four branching algorithms using the one- and three-parameter logistic models:

(1) Maximum-Likelihood Estimation (MLE)--The MLE ability estimates for the three-parameter logistic (3PL) model are obtained by the following formula:

$$\frac{\partial ln\, L}{\partial \theta_i} = \left[D \sum_{i=1}^{n} \frac{a_i(P_i - c_i)}{(1 - c_i)} \frac{v_i - P_i}{P_i} \right] \qquad [2]$$

where

$ln\ L$ = natural log likelihood function,

v_i = 1 if the item response is correct, 0 if incorrect,

D = 1.7,

n = number of items administered to examinee i, and

P_i, θ_i, a_i, c_i = as defined in Equation [1],

This branching strategy is an unbiased and relatively efficient estimation procedure. Its major drawback is that non-convergent

solutions are possible if the examinee does not get at least one item correct and incorrect (e.g., gets all administered items right).

(2) Bayesian--The Bayesian branching strategy in part addresses the non-convergence problem encountered with MLE estimates. The ability estimates are obtained by the following formula (3PL):

$$\frac{\partial ln\ F}{\partial \theta_i} = \sum_{i=1}^{n} k_i (v_i - P_i) - \theta_i \qquad [3]$$

where

$ln\ F$ = the natural log of the Bayesian posterior density function,

$k_i = \frac{Da_i(P_i - c_i)}{P_i(1 - c_i)}$

P_i, θ_i, a_i, c_i = as defined in Equation [1], and

D, v_i, P_i, and n = as defined in Equation [2]

While the Bayesian strategy addresses the problem of non-convergence, it nevertheless tends to be less efficient than the MLE solution (i.e., requires more items to estimate θ at a given level of precision) and in certain circumstances can generate biased estimates of ability.

(3) Crossover--This technique is designed to take advantage of the relative efficiency of the MLE branching strategy, yet avoids the non-convergence problem. Using the crossover technique, the examiner starts off with the Bayesian branching strategy. Once a right/wrong combination has been achieved, the algorithm switches to an MLE estimation procedure to calculate θ.

(4) Modal Bayesian--The Modal Bayesian branching algorithm is given by the following formula:

$$\frac{\partial \ln F}{\partial \theta_i} = \left[D \sum_{i=1}^{n} \frac{a_i(P_i - c_i)}{(1 - c_i)} \frac{v_i - P_i}{P_i} \right] - \frac{\theta_i - M}{S^2} \qquad [4]$$

where

M = Bayesian prior mean,

S^2 = Bayesian prior variance,

P_i, θ_i, a_i, c_i = as defined in Equation [1],

D, v_i, P_i, and n = as defined in Equation [2], and

ln F = as defined in Equation [3]

Conceptually this formula is similar to the MLE branching algorithm except that the test administrator supplies the a priori mean and variance (e.g., M = 0, S^2 = 1). Once these parameters are established, the posterior distribution will be constrained to a range that provides a convergent solution. This branching algorithm has the advantages of theoretically calculating the θ estimate in an efficient manner and eliminates the problem of non-convergence associated with the MLE solution.

Assessing Reliability

The assessment of reliability is somewhat problematic for computer adaptive testing. With traditional paper-and-pencil tests one can examine three types of reliability depending what the test is designed to measure: (1) internal consistency when the items cover some defined domain; (2) test-retest reliability if the trait is supposed to be stable; and (3) alternate forms reliability if an alternate form of the test exists. One of the assumptions underlying the classical measurement theory is that measurement precision is constant across the entire ability spectrum. This is reflected in the calculation of the standard error of measurement, the formula for which is given as

$$SEM = \sigma\sqrt{1 - \rho} \qquad [5]$$

where

σ is the standard deviation of the test, and

ρ is an estimate of reliability for the test.

With Item Response Theory, the measurement precision is thought to vary as a function of proficiency; therefore it cannot be summarized by a single overall indicator of reliability (Thissen, 1990). One could, however, calculate a marginal reliability coefficient by averaging the marginal error variance, $\overline{\sigma}^2_{e\cdot}$, across the proficiency distribution, where $\overline{\sigma}^2_{e\cdot} = \int \overline{\sigma}^2_{e\cdot}\, g(\theta)d(\theta)$. The marginal reliability coefficient is given as

$$\overline{\rho} = \frac{\sigma^2_\theta - \overline{\sigma}^2_{e\cdot}}{\sigma^2_\theta} \qquad [6]$$

where

σ^2_θ is the variance of the ability estimates.

Test-retest reliability can be achieved with CAT in the traditional way by administering a test at T_1, waiting a period of time, and re-administering the test at T_2. It should be noted that the test items for the second administration may be different from those given at T_1, but are drawn from the same item pool. However, the ability estimates of both administrations can be correlated to obtain a measure of stability.

Alternate forms reliability usually refers to the correlation of two different versions of the test across the same individuals. In CAT this would typically mean drawing items from a different item pool to avoid memory and practice effects. In this study, however, the term alternate forms is expanded to include the paper-and-pencil version of

the test. Usually correlations between paper-and-pencil and CAT administrations come under the rubric of predictive validity, but this was not applied to this study. In a predictive validity study, one uses CAT information to predict what the score will be on the paper-and-pencil test (e.g., the paper-and-pencil test is the criterion). In this study, neither version of the test was designated as the criterion. In part this was done to address the potential problem of order effects.

This study investigated both the internal consistency reliability of CAT and the alternate forms reliability of the different branching strategies when correlated with an equivalent paper-and-pencil test.

Instruments Used

Calibration and Evaluation of the Item Bank

The item bank consisted of 120 items designed to assess college mathematics ability and is used to place students in one of several mathematics course options. The test curriculum was divided into four general clusters including elementary algebra (20 items), advanced algebra (46 items), geometry and trigonometry (47 items), and calculus (7 items) (Hsu and Shermis, 1989). Because the item bank was so large, it was not possible for any *one* individual to complete *all* items. Consequently three forms of the paper-and-pencil test were constructed. Form A consisted of forty-five questions with fifteen anchor items. This form was administered to 423 students who took the exam as a practice final. As an incentive to students, extra credit was given if their scores met the minimum criteria specified by their instructors. Coefficient alpha, a measure of internal consistency, computed for this form was α=.77. Forms B and C were the same test with the items presented in different order. This test consisted of seventy-five

unique questions and fifteen anchor items, was given as part of a math competition. The highest scoring individuals received scholarship awards. A total of 937 students took the test. Coefficient alpha computed on the combined forms B/C was α=.93. The item data obtained from these tests were used to calibrate and equate the items using the LOGIST IV computer program. The steps involved in construction the test and calibrating the items are described in more detail in Shermis (1986).

Shermis (1988) and Shermis and Chang (1991) examined the assumption of unidimensionality for the math item bank. The second study used a methodology outlined by Cook, Eignor, and Dorans (1988) who advocated dividing the test up into 'parcels' and performing a second-order factor analysis. Both analyses resulted in the conclusion that the items comprising the bank were essentially unidimensional.

Methods and Techniques

Subjects.

Subjects were 68 students from two classes in a large southwestern university. One class was an introductory statistics course taken primarily by juniors or seniors from various university departments. The other class was a course in learning skills taken by students in their sophomore to senior years. Students volunteered to participate in the study. Those who chose not to participate were given alternate assignments by the course instructors. Approximately equal numbers of male and female students participated in the study.

Procedure.

The subjects were randomly divided into two groups. Subjects in group 1 took the CAT first and the paper-and-pencil test second. The

order of testing was reversed for the second group. There was a three-week time interval between the two testing sessions for each group to minimize memory and practice effects. Testing, however, was completed during the same week for both groups during each testing session. When subjects reported for testing, they were randomly assigned to take the test using one of the four branching algorithms. All subjects responded to a short attitude questionnaire after completion of their second test.

Subjects reported to a microcomputer facility on campus for their CAT session. The CAT was administered on a Macintosh II computer. A researcher demonstrated the testing procedures and made sure all students understood the test instructions. This individual also worked with the student until the examinee felt comfortable using the computer's mouse, used for selecting the response alternatives. The subjects then registered their names and identification numbers on the computer and began the testing session.

HyperCAT™ presented a test item on the computer screen. Subjects could take as much time as they desired to respond to the multiple choice item, and they were free to use a blank sheet of paper to make their calculations. They indicated their response by "clicking" on their choice. Students responded to items on the screen until the HyperCAT™ program signaled the completion of the test by indicating so on a test card. The test administrator then printed the Test Report and the Debug Field.

Testing was concluded under two circumstances. A normal test termination occurred if the ability estimate achieved an information level of 10 (or an error variance of .1 for Bayesian ability estimation). A conditional test termination occurred if (1) the current

logistic ability estimate was less than -4 or greater than +4 or (2) a maximum of twenty items had been administered. The first conditional criterion was generated to avoid non-convergence problems for the maximum likelihood branching algorithm.

The 45-item paper-and-pencil test was administered to the the subjects in each group during one testing period. This version of the test corresponded to Form A described in the instrumentation section above. To accommodate subjects who were unable to attend the group testing period, however, several other testing periods were scheduled during the same week. The subjects were given one hour for responding to the paper-and-pencil test. They marked their responses to the multiple choice questions on a machine readable answer sheet.

Results

Investigative analyses showed no initial group differences by algorithm for the paper-and-pencil test [$F(3,59) = 0.33$, $p > .05$] and for CAT-estimated ability [$F(3,64) = 1.64$, $p > .05$)]. Table 1 shows the means and standard deviations for these two variables across the four branching algorithms. The overall mean and standard deviation for the paper-and-pencil test were 17.25 and 7.30, respectively. This was substantially lower than the mean for the normative group ($z = 3.84$, $p < .001$) where the mean was $\bar{X} = 21.6$ and the standard deviation = 9.33. It is useful to note that in the sample, the mean scores of the paper-and-pencil tests roughly corresponded to the CAT theta estimates.

Insert Table 1 About Here

As a way of assessing whether order effects were present, a multivariate analysis of variance was performed. This analysis examined the means for both the paper-and-pencil and CAT-administered math tests as a function of which form was given first. The test revealed that there were no order effects [λ= .944, $F(2,60)$=1.79, $p > .05$)]. Table 2 shows the pre-posttest score and theta means and standard deviations by the order in which the tests were given.

Insert Table 2 About Here

Next, utilization of the item bank was investigated. This process involved examining two components. First, an assessment was made as to how often the items from the 120 item bank were used, and second, which algorithms, if any, used more items than others. Table 3 shows how often each item in the bank was administered during the testing periods. In all, a total of 65 unique items were used across the four branching algorithms. The average difficulty level for the 65 items was .50 with a standard deviation of .27. The weighted average (difficulty value multiplied by the frequency of use) was a more difficult .38.

Insert Table 3 About Here

In all, students averaged 14.88 items per test with a standard deviation of 5.61. This constitutes about one third the length of the paper-and-pencil form. The CAT test length was constrained, however, to be a maximum of 20 items because of time constraints. Thirty subjects were administered the maximum ceiling of twenty items. Conceivably

these individuals would have had longer tests, if the termination criterion had been set differently. Table 4 shows the breakdown of test length by branching algorithm. There was no significant difference in overall test length by this factor [$F(3,63)=2.014$, $p > .05$].

Insert Table 4 About Here

Nine individuals had tests terminated because they exceeded the lower minimum threshold of theta set at -4.0. The number of items typically administered to these individuals was less than those who had normal test terminations.

Time to complete the test was investigated as well. The average time to finish the test was 17.83 minutes with a standard deviation of 9.71. This compared favorably with the sixty minutes allocated for the paper-and-pencil form. There were no differences in time to complete as a function of CAT branching algorithm [$F(3,61)=0.474$, $p > .05$]. Table 5 lists the means and standard deviations by branching algorithm for time to completion.

Insert Table 5 About Here

A final criterion to examine how well the item bank performed had to do with computations of error variance. The mean error variance for the CAT was calculated to be .19 with a standard deviation of .19. Excluded from this analysis are the sixteen cases which had non-convergent solutions or an exceeded bound on theta. There were no differences in error variances as a function of CAT branching algorithm

[$F(3,48)=1.24$, $p > .05$]. Table 6 lists the means and standard deviations by branching algorithm for the error variance computations.

Insert Table 6 About Here

Internal consistency was computed using the marginal reliability formula in Equation 6. With a variance of theta = 6.57 and an average error variance of 0.19, the marginal reliability is computed at $\overline{\rho}$ = .97. The marginal reliabilities were broken down by branching algorithm as well. The marginal reliability for the MLE branching algorithm was lowest at $\overline{\rho}$ = .86, followed by the Bayesian algorithm at $\overline{\rho}$ = .96, the Crossover algorithm at $\overline{\rho}$ = .97, and the modal Bayesian at a very high $\overline{\rho}$ =.98.

Next, as a way to check the alternate forms reliability, correlations were calculated between the paper-and-pencil test scores and the theta estimates. The overall correlation was $r = .63$ [$n = 63$, $p < .001$]. Correlations between the CAT theta estimates and paper-and-pencil test scores was highest for the MLE algorithm [$r = .79$, $n = 17$, $p < .001$], followed by Bayesian [$r = .73$, $n = 16$, $p < .001$], Modal Bayesian [$r = .69$, $n = 13$, $p < .01$], and Crossover [$r = .48$, $n = 16$, $p > .05$], respectively[3].

Finally, as a check to ascertain whether the theta estimates could be translated into scores that would have some comparability with the paper-and-pencil version of the test, a conversion to expected scores was performed. To generate expected scores, the theta values were transformed using a function based on the test characteristic curve that

---

[3] p-values are based on Bonferonni confidence intervals.

is shown in Figure 3. The mean for the expected scores was found to be 17.59 with a standard deviation of 8.60. This is very close to the mean of 17.25 and a standard deviation of 7.30 for the paper-and-pencil version of the test reported earlier.

Insert Figure 3 About Here

Educational Importance of the Study

There are several factors that play into choosing one branching strategy over another. To convince examinees and test users of the worth of CAT, it is helpful to provide information that test scores administered adaptively on the computer and via paper-and-pencil are roughly equivalent. In this experiment, the theta estimate for the MLE branching algorithm had the highest correlation with its paper-and-pencil counterpart. The main drawback with the MLE algorithm is the requirement to have at least one item correct and one item incorrect. Unless this condition is met, an ability estimate cannot be calculated. HyperCAT™ attempts to achieve this condition by branching to the most difficult item in the bank if the first item administered is answered correctly or branching to the easiest item in the bank if the first item is answered incorrectly.

The Bayesian branching algorithm also had good alternate forms reliability. Bayesian branching algorithms have been criticized in the past as possibly being slow in reducing error variance or permitting theta estimates out of the normal (logistical) range. In this experiment, the theta estimates of the Bayesian branching algorithms were comparable to the others. The error variance was not statistically

BEST COPY AVAILABLE

significantly larger than that of the other branching algorithms and the amount of time required to administer the test was not significantly longer than the other three approaches. Moreover there is no requirement of a response pattern for a minimum of one item correct/incorrect when using the Bayesian approach.

The modal Bayesian branching strategy represents a compromise between the Bayesian and MLE approaches. It puts a bound on the range that theta can take and has been touted as coming to a relatively quick estimate of ability similar to the MLE algorithm. In this experiment, the modal Bayesian algorithm demonstrated good alternate forms reliability, but fared no better than the other approaches in reducing testing time or coming to a more precise estimate of ability.

The Crossover strategy represents a different compromise between the Bayesian and MLE approaches. It is functionally a combination of the two techniques. The computer first starts off using the Bayesian branching algorithm and then switches to the MLE algorithm once the right/wrong condition has been met. While the Crossover algorithm fared just as well as the others with respect to time of testing and error variance estimates, it had non-significant alternate forms reliability.

All four branching algorithms had good marginal reliabilities, ranging from .86 for MLE to .98 for the modal Bayesian. These reliabilities corresponded to the paper-and-pencil internal consistency measures which ranged from .77 for the 45-item Form A (used in this experiment) to .97 in Forms B/C. Items in these latter forms, along with those in Form A, comprised the math item bank.

In using the adaptive testing program, it appears as if students achieved some efficiency in the amount of time spent testing. Testing

time was about one-third of that spent on the paper-and-pencil version, though had the stop rule been more liberally set, additional testing time probably would have been required by at least half of the examinees. A savings of about one-half the time of testing is probably more reasonable when utilizing this technology. The authors are exploring the efficiency issue in a current study which compares testing times for computer adaptive tests given with the one-parameter and three-parameter logistic models.

Questionnaire data indicated that the undergraduate students who participated in this experiment had had some exposure to microcomputers and that they liked taking the test on the computer. In part this may be due to the uniqueness of the testing situation, but may attest to the fact that CAT may have a more receptive audience than was thought in the past. We hope that programs like HyperCAT™ will make it easier for measurement practitioners to create adaptive tests and begin experimenting in different subject-matter areas.

References

Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*(4), 431-444.

Cook, L.L., Dorans, N.J., & Eignor, D.R. (1988). An assessment of the dimensionality of three SAT-Verbal test editions. *Journal of Educational Statistics, 13*(1), 19-43.

Ho, R.G., & Hsu, T.C. (1989). *A comparison of three adaptive testing strategies using MicroCAT.* Paper presented at the Annual Meetings of the American Educational Research Association, San Francisco, CA.

Hsu, T.C., & Shermis, M.D. (1989). The development and evaluation of a microcomputerized adaptive placement testing system for college mathematics. *Journal of Educational Computing Research, 5*(4), 473-485.

Novick, M.R. (1969). Bayesian methods in psychological testing. Princeton, NJ.:Educational Testing Service, *Research Bulletin (RB-69-31)*.

Owen, R.J. (1969). A Bayesian approach to tailored testing. Princeton, NJ.:Educational Testing Service, *Research Bulletin (RB-69-92)*.

Reckase, M.D. (1973). *An interactive computer program for tailored testing based on the one-parameter logistic model.* Paper presented at the Nation Conference on the Use of On-line Computers in Psychology, St. Louis, MO.

Reckase, M.D. (1974). *An application of the Rasch simple logistic model to tailored testing.* Paper presented at the annual meeting of the American Educational Research Association.

Shermis, M.D. (1986). *Microcomputerized adaptive placement testing for college mathematics.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Shermis, M.D. (1988). *The use of IRT to investigate the linear/hierarchical nature of a college mathematics curriculum.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Shermis, M.D. (1991). *Computer adaptive testing: Are you ready?* Paper presented at the annual meeting of the Michigan School Testing Conference, Ann Arbor, MI.

Shermis, M.D., & Chang, S.H. (1991). *The use of IRT to investigate the linear/hierarchical nature of a college mathematics curriculum.* Unpublished manuscript, Department of Educational Psychology, The University of Texas at Austin.

Thissen, D. (1990). Reliability and measurement precision. In Wainer, H. (Ed), *Computerized adaptive testing: A primer.* Hilldale, NJ: Lawrence Erlbaum Associates.

Urry, V.W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14,* 181-196.

Table 1. The means and standard deviations for the paper-and-pencil version of the math test and the theta estimates by CAT branching algorithm (N=68).

	Paper-and-Pencil Score		CAT Theta Estimate	
Group	Mean	Std. Dev.	Mean	Std. Dev.
MLE (N=18)	18.71	7.83	0.09	1.09
Bayesian (N=19)	16.50	7.25	-0.86	2.76
Crossover (N=16)	17.19	6.39	-0.18	0.94
Modal (N=15)	16.43	8.15	-1.70	3.47

Table 2. The means and standard deviations for the paper-and-pencil version of the math test and the theta estimates by order in which the test was given (N=68).

	Paper-and-Pencil Score		CAT Theta Estimate	
Group	Mean	Std. Dev.	Mean	Std. Dev.
Pencil First (N=39)	18.44	7.00	-0.349	2.09
CAT First (N=29)	15.86	7.51	-1.65	2.96

Table 3. Frequency of administration (Freq.) and difficulty level (p) of each item used in the math item bank (N=68).

Item #	Freq.	p	Item #	Freq.	p	Item #	Freq.	p	Item #	Freq.	p
1	7	.43	31	21	.29	61	0	NA	91	4	.75
2	34	.47	32	32	.38	62	0	NA	92	0	NA
3	23	.43	33	3	.67	63	1	1.00	93	0	NA
4	5	.40	34	0	NA	64	0	NA	94	0	NA
5	24	.38	35	10	.30	65	0	NA	95	38	.53
6	0	NA	36	0	NA	66	0	NA	96	0	NA
7	34	.44	37	0	NA	67	8	.00	97	0	NA
8	0	NA	37	0	NA	68	0	NA	98	30	.60
9	15	.33	39	6	.33	69	0	NA	99	37	.84
10	15	.53	40	2	1.00	70	0	NA	100	0	NA
11	1	1.00	41	0	NA	71	0	NA	101	0	NA
12	0	NA	42	1	1.00	72	4	.25	102	21	.62
13	0	NA	43	0	NA	73	1	1.00	103	0	NA
14	0	NA	44	0	NA	74	0	NA	104	0	NA
15	0	NA	45	16	.44	75	0	NA	105	27	.19
16	13	.85	46	8	.38	76	0	NA	106	0	NA
17	25	.48	47	13	.31	77	0	NA	107	0	NA
18	1	.00	48	68	.57	78	0	NA	108	14	.29
19	4	.75	49	14	.57	79	0	NA	109	0	NA
20	10	.60	50	0	NA	80	0	NA	110	22	.55
21	25	.44	51	0	NA	81	22	.73	111	22	.73
22	28	.36	52	4	1.00	82	21	.33	112	29	.31
23	35	.23	53	0	NA	83	1	1.00	113	21	.19
24	19	.63	54	9	.56	84	0	NA	114	8	.88
25	13	.46	55	0	NA	85	5	.40	115	0	NA
26	0	NA	56	0	NA	86	0	NA	116	28	.43
27	0	NA	57	1	.00	87	4	.75	117	0	NA
28	4	.25	58	8	.75	88	0	NA	118	38	.18
29	2	.50	59	0	NA	89	1	.00	119	21	.57
30	15	.40	60	2	.50	90	2	.00	120	36	.42

Table 4. The means and standard deviations for the number of items administered by CAT branching algorithm (N=67).

	Number of CAT Items	
Group	Mean	Std. Dev.
MLE (N=18)	16.56	4.22
Bayesian (N=19)	15.05	5.59
Crossover (N=16)	15.87	5.66
Modal (N=14)	12.07	6.13

Table 5. The means and standard deviations for time to complete the test by CAT branching algorithm (N=66).

	Time to Complete the CAT (in minutes)	
Group	Mean	Std. Dev.
MLE (N=17)	20.41	10.28
Bayesian (N=19)	17.00	8.47
Crossover (N=16)	17.87	8.40
Modal (N=14)	16.93	11.68

Table 6. The means and standard deviations for error variances by CAT branching algorithm (N=66).

	Error Variances	
Group	Mean	Std. Dev.
MLE (N=15)	0.164	0.078
Bayesian (N=16)	0.268	0.310
Crossover (N=10)	0.154	0.062
Modal (N=11)	0.153	0.137

Figure 1. An example of what a test item might look like in HyperCard™.

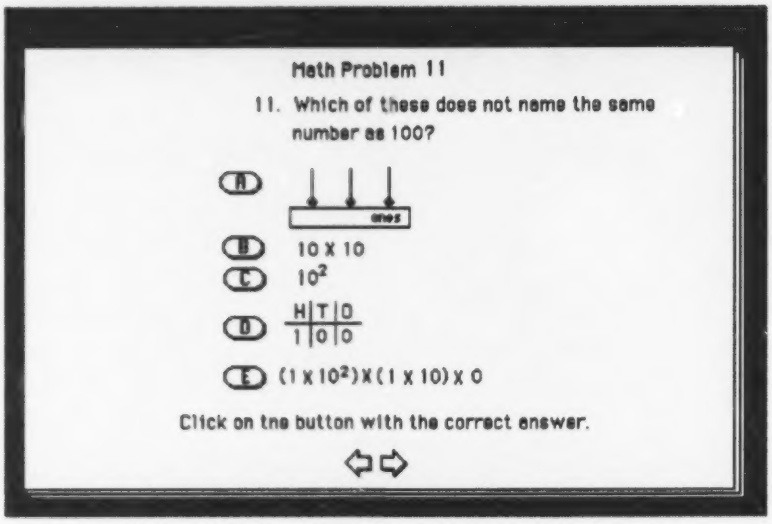

Figure 2. An illustration of the HyperCAT™ Card Manager.

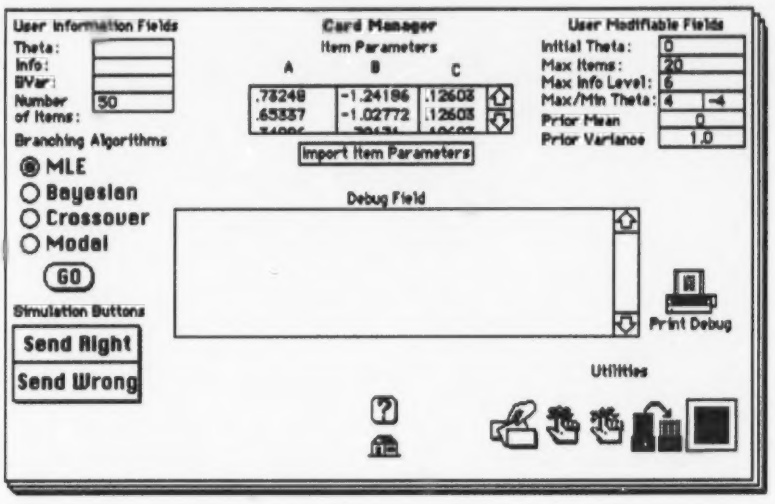

Figure 3. The Test Characteristic Curve (TCC) for the math item bank.